DOCUMENT RESUME

ED 168 913 SO 011 397

AUTHOR Shaver, James P.
TITLE Design Considerations for Classroom Research.
PUB DATE 23 Nov 78
NOTE 15p.; Paper prepared for Annual Meeting of the National Council for the Social Studies (Houston, Texas, November 23, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Research; Data Analysis; Decision Making; Educational Improvement; Educational Objectives; *Educational Research; Elementary Secondary Education; Information Needs; Predictor Variables; *Research Design; Research Methodology; Research Needs; *Research Utilization; Statistical Analysis; *Teacher Role

ABSTRACT

Classroom teachers need to make sound judgments and decisions concerning curricular and instructional issues. The teachers who wish to become more effective in the classroom should learn to develop their own research designs since educational research reported in journals is often inconclusive, conflicting, or not relevant to the classroom teachers. Experimental studies can be particularly helpful to teachers if they investigate the effects of variables such as textbooks, tests, and homework assignments on one or more other variables such as student knowledge, student attitudes, and length of time required to complete a test or homework assignment. Factors to be considered when designing an experimental study include experimental validity (comparing and contrasting results with results from a control group or situation), internal validity (extent to which the observed effect appears to be due to one's experimentation), and external validity (determination of persons and circumstances to which the results apply). The conclusion is that teachers can do valid classroom research if they combine simple descriptive statistics with common sense and exercise care in maintaining experimental validity when they gather information. (DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED168913

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James P. Shaver

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

DESIGN CONSIDERATIONS FOR CLASSROOM RESEARCH*

James P. Shaver
Utah State University
Logan, Utah

*Paper prepared for a section meeting, "Teacher Research in the Classroom: How to Do It", annual meeting of the National Council for the Social Studies, Houston, November 23, 1978.

SO-011397

JAN 19 1979

DESIGN CONSIDERATIONS FOR CLASSROOM RESEARCH*

James P. Shaver
Utah State University
Logan, Utah

Why should a classroom teacher do research? Realistically, not with the intent of making a contribution to Scientific (with a capital S) knowledge--the systematic explanations of phenomena that are labeled "theory". Teachers' other professional interests legitimately consume so much of their time and energy that little of either is left over for conceptualizing scientific studies and seeking the resources to carry them out. However, teachers are instructional decision-makers for whom systematic data can be of considerable assistance. Much relevant information is not available, unless teachers gather it through their own efforts. The findings of educational research reported in journals are too often inconclusive, conflicting, or not pertinent to the matters that concern teachers either in building instructional programs or in interacting with students day-by-day. (The lack of fruitfulness of research in social studies education has been most recently documented by Wiley, 1977.) Teachers who want data as a basis for decision-making will have to produce much of it themselves. To do so scientifically (with a lower case s), systematically and objectively, is possible, even within the constraints of the teaching situation.

Variations of experimental studies--i.e., studies in which some variable (e.g., the textbook, the quantity or quality of resource materials, types of film, types of homework assignments, types of items on tests) is investigated in order to determine its effect on one or more other variables[1] (e.g., student knowledge, student attitudes, proportion of completed homework, length of time to complete a test)--hold the most promise for teachers who are interested in improving their instructional effectiveness. In designing such studies, there are some common sense notions--educational researchers talk about them using rather technical language--which can be helpful in producing valid results.

Experimental Validity[2]

Experimentation as a means of gathering information depends on, among other

---

*Paper prepared for a section meeting, "Teacher Research in the Classroom: How to Do It", annual meeting of the National Council for the Social Studies, Houston, November 23, 1978.

[1]The first type of variable is typically called an independent variable, or a treatment variable. The second is called a dependent variable, because the researcher wants to know if values on it (e.g., number of correct answers on a test) are dependent upon the treatment variable.

[2]The following discussion relies heavily on a classic analysis of research by Donald T. Campbell and Julian C. Stanley, first published in the Handbook of research on teaching (edited by N.L. Gage and published by Rand McNally, 1963) as Chap. 4, "Experimental and quasi-experimental designs for research on teaching". The chapter has been reprinted as a separate paperback book by Rand McNally with the shorter title, Experimental and quasi-experimental designs for research.

things, being able to make comparisons and contrasts. Imagine, for example, a teacher teaching an "Energy and the Environment" Unit for the first time. She wants to know whether student knowledge about alternative sources of fuel is greater as a result of the unit. She teaches the unit to one of her classes and gives them a final exam. The teacher has used what is often referred to as the one-group, posttest design. She knows the students' scores on the final exam; but, without some basis for comparison, she does not know if her students' scores are any, or much, different than they would have been without the unit. (She may, of course, make an "intuitive" comparison--such as to the information indicated by the students' comments prior to studying the unit. Such observations are important sources of information, but they are fraught with opportunities for invalid conclusions.)

To improve her design, the teacher might decide to obtain another set of scores from her students to compare against the final exam scores. For example, she could administer a test to her students before the unit and then compare scores on the final exam to those pretest scores. This would be what is termed a one-group, pretest-posttest design. It is somewhat better than the one-group, posttest design, but not much. To understand why, it is helpful to consider some of those common sense notions that educational researchers have about designing studies to get valid results--i.e., so as to have experimental validity.

Internal Validity

Experimental validity is commonly considered to have two aspects--internal validity and external validity. Internal validity has to do with the extent to which you can assume that any observed effect (e.g., gains on the test of knowledge of alternative fuel sources) is due to your treatment; or, put conversely, the extent to which you can assume that the treatment's effect has been observed (it might be, for example, that students learned from the unit on energy, but for one reason or another the learning was not reflected in the test scores.) External validity has to do with the extent to which you can generalize your finding(s) beyond the particular study from which they were obtained. (E.g., if the students gained in their knowledge of fuel sources, can the teacher assume that she will attain the same results using the unit with future classes?) Internal validity is the more important of the two, because unless one can be fairly certain that his or her treatment has had the desired effect, it is meaningless to ask with whom or under what conditions the effect can be expected to occur.

Threats to Internal Validity. Being aware of several common threats to internal validity can help teachers to design studies that will provide more reliable information for their decision-making purposes. The one-group, pretest-posttest design, mentioned above, is subject to most of the threats that need to be considered. Let us assume that the teacher using the Energy and Environment Unit obtained as large a gain in scores from the pretest to the posttest as she had hoped for. What might account for the gain other than the treatment (the unit)?

One threat to the internal validity of her result is testing. It may be that taking the pretest affected the students in some way (they learned

the names of various fuels from reading the multiple-choice items, or taking the test alerted them to news items they wouldn't have noticed otherwise, or the test piqued their interest so that they sought readings about fuels outside of class), so that what looked like a gain due to the unit was really due to taking the test. Also, if a test is not valid--i.e., if it does not measure the teacher's instructional goals, treatment effectiveness, or lack of it, may not be detected.

Another threat to internal validity is what researchers call history. This term is used technically to refer to experiences other than the treatment variable that the students might have between the time a treatment starts and the time it ends. (Social studies teachers are used to thinking of history as what has happened in the past, e.g., what had happened to the students before a new unit is taught. Those prior experiences may be a threat to validity, but researchers talk about that under the term "selection", which we will discuss shortly.) For example, many of the students in the Energy Unit teacher's class may have watched a TV special on energy during the duration of the unit, with that experience accounting for their better scores on the final exam.

Another possible threat to internal validity is instrument decay. That is, changes in the test (e.g., if different pretest and posttests were used) or in scoring the test might produce a change in scores. The latter source of instrument decay is a particularly likely threat. If a teacher knows which are the pretests and which are the posttests, she or he may, consciously or unconsciously tend to score them differently. For example, an expectation that students will do better on the posttest might affect one's scoring judgments. Scoring "blind" (without knowing which are the pre- and which the posttests) is a good idea.

Other threats to internal validity are not likely to have affected the results of the Energy Unit Research. Nevertheless, they bear mention because of their applicability to other research studies that teachers might want to do.

One such threat is maturation. That is, sometimes an observed effect is due to changes in the students that occur as a function of the passing of time. For example, if Piaget is correct, we can expect children to move from the preoperational stage of thinking to the concrete operational stage at about age seven. Imagine a teacher who throughout the year uses a set of exercises with her second grade class, hoping to increase the students' ability to think in concrete operational terms. To determine effectiveness, she uses the one-group, pretest-posttest design. She finds a large average gain in scores; but, the gain might be due only to normal maturation rather than to her exercises.

Maturation can have a deleterious effect, too. Fatigue or hunger are considered maturation processes. If, for example, the Energy and

Environment Unit was taught in the late afternoon when students were fatigued, that fatigue might counter any positive effects of the unit.

Another threat to internal validity is that of statistical regression. This threat has a rather complex statistical explanation, but it also can be explained in a fairly common sense way. Statistically, we say that students who had extreme scores on a pretest will have scores closer to the group mean (the arithmetic average) on the posttest, even without treatment. Let us say, for example, that our Energy Unit teacher is especially interested in helping "slow" students do better. So she administers her pretest and later selects for analysis those students who did most poorly on it (perhaps those who had the bottom ten percent of the scores). She compares the mean pretest score of this selected group with its mean posttest score. She would likely find a gain, because we would expect the scores of the students to move toward the group mean--i.e., to be higher in this case. That expectation can, as I mentioned, be explained statistically. But that involves getting into such matters as normal probability distributions. On a more common sense level, we can think of the students who got the lowest ten percent of scores as probably knowing less than many of the other students, but also as likely to have had "bad luck" on the first taking of the test--they guessed poorly or happened to be especially fatigued or emotionally upset for some reason. On the posttest, they are likely to have better "luck" while other students have "bad luck". So, while the selected group of students will still have low scores, their scores will tend to be somewhat better than on the pretest (even if they had not been exposed to the energy unit). They will have moved toward the mean--and other "low knowledge" students who had "bad luck" on the second testing will have even lower scores.

Note that the regression effect works at both ends of the distribution. If the teacher had picked the ten percent of students with the highest scores on the pretest, their posttest scores would likely have gone down. She might have concluded that the treatment was not effective with "bright" students. But the notion of "good luck" is as applicable to students who would do well anyway as the notion of "bad luck" is to students who do poorly. Just as some students would be in the bottom ten percent because of "bad luck", some would be in the top ten percent because of "good luck", and their scores would be likely to move toward the mean on the posttest.

This is, of course, a much over-simplified discussion of statistical regression. The major message is, however: Be careful when comparing pre- and posttest scores for students selected because of extreme pretest (or other, such as IQ or social adjustment) scores. It may appear that your treatment had an effect when there was none; or an effect that did occur may be obscured.

Two More Threats and Design Considerations. What to do about these threats? The researcher's answer has been to add one or more comparison groups to the design. These are often called control groups, although that can be a misnomer, as it suggests that nothing happens to them, when in reality they usually receive some alternative treatment.

One "comparison group" design that is used on occasion provides an opportunity to discuss two more threats to internal validity. It is called the static group design. In this design, two "natural" groups are compared--one of which has had the treatment of interest, the other of which has not--but with no opportunity to administer a pretest. For example, the end-of-the-year standardized achievement scores of a group of students who were part of a class that included a political participation project might be compared against the scores of other students in the school who took a social studies course without such a project.

Lack of control over how students got in the project class might result in differential selection, a threat to internal validity. If, for example, students were free to choose which class they enrolled in, their reasons for enrolling in the participation class or the other might be based on factors related to achievement test performance, such as interest in social studies. There would be no way to establish that the groups are equivalent, for in whatever other ways they may be similar, the students are different in one crucial regard--one group signed up for the project class, the other didn't. Effects might be due to the treatment (the participation project) or to the initial differences in the groups.

Sometimes experimental mortality is a threat to internal validity, too. Here mortality is not used in the sense of students or teachers dying, but in the sense that there may be a differential loss of students from the groups compared. For example, if students did not know about the political participation project beforehand, and were allowed to drop the class if they wished after finding out about it, that would be experimental mortality. Just as with the threat of selection, mortality means that the groups may be different in important ways (the importance depends, of course, on how many students drop out and on how different they are from those who stay) which are often nearly impossible to determine with any precision. What appears to be a difference due to treatment may simply be the result of losing students.

Where possible, researchers would like to select their own treatment and control (or alternative treatment) groups. In particular, the best procedure is one that ensures that students are randomly selected to the groups. This could be done by pulling names out of a hat, as well as by the statistically sophisticated use of a table of random numbers. Any procedure that ensures that each person has an equal chance of being chosen for each of the groups is satisfactory. Random selection guarantees that there will be only chance differences between the groups.

Of course, once in a while, chance differences between groups will be large. For example, even though there are twenty girls and twenty boys in the group from which an experimental and a control (or alternative treatment) group are chosen and a random process is followed, by chance a large proportion of the girls might end up on one group and a large proportion of boys in the other.

Stratification on a characteristic such as sex can be helpful in ensuring that it will be properly represented in both groups. That is, the girls and boys can be treated like two different groups (i.e., stratified) for purposes of selection, with the random selection process applied to one group first and then to the other. Multi-strata can be used. For example, prior to selection for a Energy Unit study, the groups of boys and girls could be further stratified according to whether they had previously taken a biology course or not.

Matching can also be an aid at times. To ensure that all IQ levels are represented in both groups, a teacher could rank all of the students according to their IQ scores, put into pairs those with the closest rankings, and then randomly assign (perhaps by flipping a coin) the members of each pair to the experimental or control group. One might also want to match on other variables, such as GPA. But when two or more matching variables are used, problems often occur because good fitting pairs cannot be found. There is always the student with an IQ of 140 and a D- GPA, or an A GPA and an IQ of 90, who has no counterpart.

Also, of course, matching can be done within strata (e.g., match boys on IQ, and then randomly select the groups; and do the same for the girls.) When only one or two stratification variables and one or two matching variables are used, the procedure is not too cumbersome; but it can easily get out of hand.

If random selection of students for the experimental and control (or alternative treatment) groups is possible, a major step has been taken toward controlling many of the threats to internal validity, because it can be assumed that there are only chance differences between the groups. (If stratification and/or matching can be used sensibly, all the better.) For example, the threat of maturation is neatly controlled by randomization, as is differential selection. So is statistical regression, if the students with extreme scores are randomly selected into the experimental and control groups. Then the tendency for the person's scores to move toward the mean on the second testing will be expected to operate equally (within chance differences) on both groups, and any change for the experimental group above and beyond that of the control group may be attributed to the treatment--if other threats to internal validity have been controlled. The same is true for testing. Any effect of the pretest should be present for both the experimental and control groups, so any gain by the experimental group over the control groups may be attributed to the treatment--if no other threat accounts for it.

History and mortality are not controlled so neatly. For example, the location of a classroom may cause a history effect not controlled by random selection--perhaps a positive effect if right next to the school media center; a negative one if right next to a classroom in which another teacher has trouble controlling the students. And the means of selection will not keep students from dropping out, especially from voluntary programs. (Sometimes rate of dropouts is a relevant dependent variable.) If the dropout rate from an experimental or control group seems high, it is a good idea to check the characteristics of those dropping--e.g., pretest scores, GPA, reading scores,

sex--to see if they differ from the students who stay and thus might have an effect on your results.

Of course, while random assignment is a major assist in securing valid research results, and therefore, worth emphasizing--teachers often cannot randomly assign students for their research projects. Nevertheless, instances where it is possible to do so should not be overlooked. For example, in a team teaching setting we were once able to assign students randomly to small group sessions in which different discussion styles were used. And in studies involving the manipulation of materials with students unaware of the differences in materials, randomization is sometimes readily accomplished. For example, if a teacher wanted to know whether putting the essay items first on a test containing essay and objective items made a difference in student performance or student attitudes toward the test, the test could be made up in the two formats and handed out to students on a random basis. Such a design is difficult to use if students can observe each other's materials. The teacher needs to be ready with an explanation when Johnny cries out, "But Billy has a different test than I do!"

Even without random selection, the use of a comparison group can be helpful in interpreting your research results. With awareness of the potential threats to internal validity, use your good judgment to obtain a control group that is as similar as possible to your treatment group. And, even if you cannot select individual students, you may be able to decide randomly--say, by the flip of a coin--which of the two groups will be given the special treatment that is to be evaluated. Then gather any indications of initial group differences--again, such as pretest scores, GPA, sex, reading scores--and take these into account in weighing your results.

There are statistical techniques for adjusting group posttest means to take into account initial differences, such as on the pretest. But teachers doing research will often not know about such techniques or have the facilities or the time to do the computations. That need not be a serious disadvantage. In fact, not relying on statistical analysis can be an advantage in that it forces you to examine your data. You should specify ahead of time how much of a gain by your treatment group over your control group would satisfy you that the treatment was sufficiently effective to be continued. You probably would want to anticipate a larger gain if the new treatment was costly--in money or your time--than if not. Then, first, compare the treatment group's pretest scores with its posttest scores (to make certain a gain occurred); second, compare the treatment group posttest scores with those for your control group to determine that the difference reaches your criterion. Then, <u>interpret</u> the result <u>carefully</u>, taking into account any potential threats to internal validity--especially differential selection.

And, whenever possible <u>replicate</u> your study. That is, do it again--on different groups, in different semesters--to see if the same results occur. The more times they do, the more confidence you can feel in the treatment.

<u>Other Designs</u>. To this point, I have emphasized the use of two or more groups in order to have a basis for comparison to determine if your treatment

did have an effect. There are single group designs that can be valid and useful for the teacher, because they allow the demonstration of effects. One of these designs is very powerful if you are concerned with behavior that is repetitive and can come and go--such as disruptive classroom behavior--rather than learnings that are more lasting (i.e., not readily subject to reversal--such as being able to explain the functions of separation of powers in our governmental system). This design is often called the ABA design. With it, one first obtains an estimate of the behavior to be changed. This might involve counting the number of times that students are out of their seats during several class periods. These pretreatment data are called the baseline. It is the base for comparison. Next, the treatment is introduced (e.g., allowing students to talk to a buddy for five minutes at the beginning of the next class period if they stay in their seat for a specified period of time) and out-of-seat behavior is counted again. If the frequency goes down, you may assume the treatment had an effect. To provide a further test, the treatment is removed, and the number of times that students are out of their seats is counted again. If the out-of-seat instances go up, then there is strong evidence that it was the treatment keeping them in their seats during the experimental period. There are, then, three phases in the ABA design--the baseline phase (the first A), the treatment phase (the B) and the measurement phase following withdrawal of treatment (the second A).

An alternative design is available in cases where the students might react to withdrawal of the treatment ("How come we aren't getting to talk for staying in our seats like we did last week?") or the outcome of interest wouldn't be expected to change as the result of withdrawing the treatment (one wouldn't expect students who learned to explain separation of powers through a special reinforcement program to forget the explanation when the reinforcement was removed). This design is called multiple-baseline design. Again, baseline data are collected, but the treatment is introduced to different students or groups of students at different times to see if change occurs with introduction of the treatment. This design could be used when, for example, the teacher had two or more classes, all studying the same subject area.

The above designs are variations of what is termed the time series design. In a time series study, the dependent variable is assessed at different points in time prior to the treatment. Then the treatment is introduced, and more measurements of the dependent variable are obtained. The series of measurements (often the means of the various assessments) is studied to determine if there was a change in pattern following the treatment. (The nature of the expected change should be predicted beforehand as a basis for demonstrating that the treatment had an anticipated effect.) The study to determine a way to keep students in their seats would fit this design well, perhaps better than the ABA design. Counts of out-of-seat behavior could be taken on several consecutive days, the treatment introduced, and counts of instances of out-of-seat behavior continued for several more days while the treatment continued. If out-of-seat behavior went down and stayed down as predicted, following the introduction of the treatment, this would be powerful evidence for the effectiveness of the treatment. Of course, it would be important to check behavior for a sufficiently long period of time to ensure that the result was not a transient one, going away, for example, when the newness of the treatment wore off. Replication with other groups is important with time series studies because this design is especially vulnerable to the threats of history. Could

something else, such as a stern reprimand and threat of punishment by the principal, have caused the change? Instrument decay must also be guarded against. For example, over the period of time could the teacher simply have become careless about counting times out of seat?

Some Comments. It is not likely that you will be able to control all of the potential threats to internal validity in your classroom research. But then, educational researchers can rarely do so in their studies either, especially when they are working in applied areas. By being aware of the threats, however, you can make some design decisions to help avoid them. And such awareness can also help you in interpreting your findings. Your knowledge of your students, your school, and your community will be invaluable as you decide if any of the threats may have contaminated your results. You will probably want to be circumspect in drawing conclusions from your results if they have not been replicated on more than one group and for more than one unit or semester. Such replication is important not only for building your confidence in whatever treatment effects you have observed, but in deciding, if they came out as you wished, how generalizable they are. That takes this discussion to the other aspect of experimental validity--external validity.

External Validity

The basic question of external validity is, To what persons and to what circumstances do your results apply?[3] The answer to this question requires, first of all, a careful, common sense look at the students in your research group(s). Are they like the other students with whom you would like to use the materials, teaching method, or whatever you are trying out? (I.e., do they represent the population of interest to you?) Is there anything about the students in your treatment group that might make the materials, etc. work especially well or poorly? Or, is your control group such (e.g., poorly motivated) that it makes your treatment effect appear greater than it is?[4] An excellent way to answer these questions, aside from your own best judgment, is to replicate your study. Repeat it with different groups, especially from one school year to the next.

A critical aspect of external validity is you, the teacher. If you are not interested in advocating that other teachers use your experimental treatment, your problems of generalization are simplified. You will not have to worry about how representative you are of other teachers. But the external validity of your results as they apply to your future use of the treatment have to do with the way in which you handled the independent variable. You should try to be certain that you are conscious of the way in which you administered the treatment. If, for example, you are interested in the extent to which different types of homework assignments result in

---

[3]External validity is discussed in Campbell and Stanley (1963). A more extended treatment is available in Bracht and Glass (1968). Also, many educational research textbooks will discuss both internal and external validity in greater detail than I could in this paper.

[4]Such an instance is reported in Oliver and Shaver (1974, Appendix, Sections 2 and 3).

students completing their work on time, you need to be certain about the important dimensions of assignment giving, so that you can later do so in the same way. For instance, were the assignments given orally, in a mimeographed handout, or written on the blackboard; at the beginning or the end of class, etc.? Basically, the description of the independent variable is critical so that the effects of its use can be anticipated validly in future classroom use, and/or so that it can be replicated for further research that might be desired.

Other threats to external validity have to do more directly with the environment you establish, consciously or not, for your research. For example, if you tell your students that they are part of a piece of research you are doing, this could lead to several threats to external and internal validity. One is the Hawthorne Effect. That is, your students may behave differently because they know they are part of an experiment. The counter to the Hawthorne Effect is the John Henry Effect--students who know they are in a control group may work harder to do well because they are not going to be shown up by the experimental students. There also is the experimenter effect. You may convey your expectations to the students in a way that influences their behavior. Any of these three effects may produce a change in the students that is mistaken for the effect of the treatment (internal validity). Because these effects are less likely to occur in future use of your experimental treatment, they are threats to external validity.

The solution to the Hawthorne and John Henry effects is either to conceal from the students that they are part of a research project or to build that impression into future uses of the treatment. The latter might involve trying to capitalize on the Hawthorne effect by becoming known as an innovative, experimental teacher.

Related to the Hawthorne effect is the novelty and disruption effect. If your treatment is a new, novel experience for the students, or if it upsets the usual classroom routine, that may affect your results. You may not be able to generalize to later classes you teach for whom the treatment has become commonplace.

You also need to be sensitive to multiple treatment effects. These are effects produced by exposure of students to two or more treatments. To go back to the Energy and Environment Unit: If the teacher had just completed with the students an experimental unit on "Population and Starvation", it could be that positive results that seem due to the Energy unit are the result of the combined effects of the two units. She may be able to generalize only to situations in which students study both units. In a sense, this becomes a question of selection (i.e., to what population can she generalize?)--or, put differently, of a selection by treatment interaction. That is, there was a combined effect of prior experience and the unit. This sort of interaction might occur for other reasons--e.g., because the experimental students were especially able academically or had other characteristics that made the treatment more effective. Replicating a research study with groups of students who have differing characteristics, such as you might encounter in your classes, helps to establish generalizability. You might also want to look at the results for different subgroups within the experimental class(es)

to see if there was an interaction effect. For example, did boys and girls learn equally well with the Energy Unit? (Boys' concerns with cars might make them more interested in potential fuel scarcities, for example.)

Similarly, history and treatment may interact. That is, the Energy Unit might be effective only because of current media attention to an energy crisis. The same effectiveness could not be expected without such media "assistance."

Testing is also very important to external validity, as it is to internal validity. One aspect of testing and generalizability is the need to be careful about expecting the same results with a test or tests different from the test or tests used as dependent variables. Just because your students did well, for instance, on one test of critical thinking doesn't necessarily mean they will do well on another. Also, testing may interact with the treatment. Taking a pretest may sensitize students to the content of an Energy Unit so that they learn more than if they had not been pretested. This is a potential threat only if a pretest is not always given with the unit. The post-test may also provide a "learning" effect, but this is rarely a problem in classroom research as testing following a set of learning activities is common. The time of testing may, however be important. How well students do on a test may depend on whether it's given right at the end of a unit or two weeks or six months later.

Most of these threats to external validity can be minimized. Common sense solutions involve such things as not letting the students know they are part of a research project (this can raise ethical problems if the content is experimental and possibly objectionable to some parents, or if participation in your project might keep students from learning things expected of them by parents, other teachers, or school district requirements) or, if it is not possible to disguise the use of new materials, by also using "new appearing" materials in your control group(s) if any are used. Again, replication is vital to determining if you will obtain the same results with groups of students with different characteristics, but similar to those students you might teach, or at different points in time or after continued use. As mentioned above, looking at subgroups of students can also be helpful (but be careful of the regression effect) in determining how generalizable your results are.

Statistics

Do you need to know statistics to do valid classroom research? No, you do not. It may be helpful to be able to compute some simple descriptive statistics: measures of the central tendency of scores for your group--such as the mean (the arithmetic average), the median (the point above and below which fifty percent of the scores fall), or the mode (the most frequently occurring score)--or of the dispersion or spread of scores--such as the range (the highest minus the lowest score plus 1) and the standard deviation (the mean squared deviation about the mean--a somewhat more complicated statistic described in every elementary educational statistics book).

Measures of dispersion are particularly important in determining if

your treatment is effective for all students. You may find that whether the central tendency of the scores changes or not, the dispersion has, because some students do particularly well with your treatment and/or others do particularly poorly with it.

If you do use measures of central tendency and dispersion, be wary that you do not depend on them too heavily. Reliance on descriptive statistics can obscure many interesting insights into what has happened to your class as a result of the treatment. Still inspect your data (for example, examining individual tests) and use your wealth of knowledge about your students, your school, and your teaching to interpret the results. You may find yourself talking to students to discover answers to questions raised by your inspection of the data. (E.g., Why did the girls do better or more poorly on an Energy Unit?) This is an important data-gathering technique--one that educational researchers often feel uncomfortable using because they have been educated to be concerned about maintaining a formal design and data-collecting techniques.

Well, how about inferential statistics, such as analysis of variance? These are of little use for classroom research such as discussed in this paper. Techniques such as analysis of variance are used to determine whether your results might have occurred by chance if your groups had been drawn randomly from the same population. Not only will teachers doing classroom research rarely have the chance to select their groups randomly, but they are not likely to be interested in generalizing to broad populations as educational researchers are (but who, alas!, often also lack randomly selected groups). A better bet for the teacher is to specify ahead of time what changes will be educationally meaningful (e.g., How many more of my students must hand in their homework before I adopt the new method of giving homework assignments?) and then check your results against that criterion. Educational significance is much more important in the classroom setting than statistical significance. And using replication to establish that the results can be attained again is more powerful than statistical analysis, too.

If you do know about inferential statistics, especially nonparametric ones (ones that make no assumptions about the populations from which your groups are drawn) such as chi-square, don't hesitate to use them. You may want to ask, for example, how likely it is that a particular distribution of scores could have occurred by chance. Or analysis of covariance can be of some assistance by making statistical adjustments for initial differences between treatment and control groups. But don't become over-reliant on inferential statistics so that questions of educational significance are overlooked, or so that you don't trust your own insights into what happened.

Remember, too, that the inferential statistics model is basically a yes-no, decision-making one--i.e., Can the result be accepted as non-chance or not? Teachers will probably more often be doing research from a developmental model. They will often be asking questions such as, "How can I improve this unit?", not "Should I teach this unit at all?" Inferential statistics are not much help with the former type of question.

Conclusion

Using your own intellectual resources to examine your data, to contemplate what went on during the treatment, and to interpret the results, and using informal ways of determining how and why your students reacted as they did, are critical. The threats to experimental validity discussed in this paper may help you to be aware of possible errors and take them into account in drawing conclusions about a treatment's effectiveness and the extent to which it is generalizable to other classes you will teach. The discussion of threats is meant, as is the discussion of designs, as an aid to teachers concerned with making sound judgments about the curricular and instructional issues that concern them.

REFERENCES

Bracht, Glenn H., and Gene V Glass. The external validity of experiments. American Education Research Journal, 1968, 5 (No. 4), 437-74.

Campbell, Donald T., and Julian C. Stanley. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963. Also, Ch. 4 in N.L. Gage (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Oliver, Donald W. and James P. Shaver. Teaching public issues in the high school. Logan, Utah: Utah State University Press, 1974 (first published by Houghton Mifflin, 1966)

Wiley, Karen B. The status of pre-college science, mathematics, and social science education: 1955-1975. Vol. III: Social science education. Report to the National Science Foundation. Social Science Education Consortium, Inc., 1977. Washington, D.C.: Government Printing Office, #038-000-00363-1.